If you study a word do you use it more often? Lexical repetition priming in a corpus of Natural Semantic Metalanguage publications

Karen Sullivan

Abstract

Psycholinguistic and corpus studies have identified syntactic repetition priming as an influence on linguistic analyses. The impact of lexical repetition priming on these analyses, on the other hand, has not been assessed. The current study finds evidence of lexical priming in a corpus of linguistics publications on the Natural Semantic Metalanguage (NSM), in which semantic analyses are written using several dozen ‘semantic primitives’ such as something, know and place. NSM theorists are repeatedly exposed to a small set of words, much like subjects in lexical repetition priming experiments. When all analyses written in NSM are removed from NSM publications, these texts are found to nevertheless include significantly more ‘primitives’ than control publications, suggesting that the study of particular words can affect linguists’ lexical choices.

Keywords: Corpus linguistics, introspection, lexical priming, Natural Semantic Metalanguage, repetition priming.

1. Introduction

Imagine that you are writing an article on the word obtain. As you write the paper, you may notice yourself writing that facts ‘obtain’ in a particular situation, or describing the difficulties that ‘obtain’ with ‘obtaining’ those facts, and generally using the word obtain more often than you otherwise might. When you have recently encountered a particular word, you recognise it more quickly and are more likely to produce it in your speech or writing. This phenomenon is called lexical repetition priming (Forbach et al., 1974;
Forster and Davis, 1984; Thomas and LaBar, 2005; and Kaschak, 2007). The related phenomenon of syntactic repetition priming has been studied by psychologists and corpus linguists (Gries, 2005; Schütze, 1996; and Szmrecsanyi, 2005), and has been found to influence linguists’ syntactic judgements (Nagata 1987a, 1987b, 1988, 1989; Schütze, 1996; Snyder, 1994; and Spencer, 1973). However, lexical repetition priming has not been studied in corpora, and the influence of lexical repetition priming on linguists has not been considered. This study offers a starting point to address these issues by examining lexical repetition priming in a corpus of linguistics publications.

Linguistics publications were selected for three reasons. First, the identification of lexical repetition priming is simplified in linguistics texts. Lexical repetition is potentially more problematic to examine in a corpus than syntactic priming, given that semantically comparable syntactic alternations (such as the dative alternation) have no direct counterpart in the lexicon. Semantically comparable syntactic constructions allow researchers to measure the extent to which one of two comparable constructions primes the choice of the next (their ‘switch rate’; see Gries, 2005). This method cannot so easily be employed to assess lexical priming, given that no comparable ‘alternation’ between near-synonyms has been demonstrated. However, lexical priming can be studied in a corpus of texts by authors exposed to a specific set of lexical items, and compared to a control in which the authors were not exposed to that set of words. Linguistic texts that study specific lexical items are, therefore, well-suited to a corpus analysis of lexical repetition priming, after instances of mentioning, citing and quoting the words under examination have been removed from consideration.

The second reason that I selected linguistic texts for this study is that linguists’ interaction with words may render them especially prone to repetition priming. Studies suggest that when you perform tasks with a particular word, the word affects you to a greater degree and results in stronger priming effects (Becker et al., 1997; and Joordens and Becker, 1997). Linguists, of course, are intensively engaging with the words they analyse, and continue to contemplate these words as they describe their findings—for example, in linguistics publications. In addition, emotional arousal associated with a particular word increases repetition priming effects (Thomas and LaBar, 2005). Linguists presumably have stronger emotions associated with their own topics of study than with arbitrarily chosen words, and this emotional arousal could contribute to repetition priming associated with these topics.

The third reason to choose linguistic texts is the special relevance that lexical priming may hold for linguistic theory. The study of syntactic priming has drawn attention to the fallibilities of introspective grammaticality judgements in syntactic theory, and encouraged the use of methodological safeguards or alternative data sources such as corpora (Nagata, 1988; Schütze, 1996; and Spencer, 1973). However, introspection remains a common method in many fields of linguistics, including those that are more concerned with the lexicon than with syntax. If lexical semanticists, for
instance, are affected by the words they study, their intuitions may not be the same as those of a representative speaker of the language. An increased awareness of priming effects might, therefore, encourage lexical semanticists to take measures to avoid priming effects or to seek non-introspective data sources.

This study focusses specifically on lexical priming by linguists working within the Natural Semantic Metalanguage (NSM). NSM theory proposes that all semantic description can be accomplished with sixty-five ‘semantic primitives’, represented by English words such as something, good and want (Goddard, 2008). NSM analyses, when conducted in English, consist of the repeated use of these few dozen English words. To the extent that NSM researchers are repeatedly exposed to a small set of words, NSM research resembles laboratory experiments on repetition priming. Even critics of NSM must contemplate the system of semantic primitives in order to formulate their criticisms. For this reason, papers supporting and opposing NSM, paired with control papers by the same researchers, were selected as the starting point for the study of lexical repetition priming in corpora and in linguistic research.

2. Method

This study involves a quantitative examination of comparable corpora. Sets of matched corpora are frequently employed to compare two or more languages (Zanettin, 1998), translated and non-translated texts (Baker, 1996; and Olohan, 2004), or texts from native speakers and language learners (Ringbom, 1998; and Valenzuela and Rojo, 2008). In this case, the two corpora compare semantics texts that mention NSM and semantics texts that do not.

Two comparable corpora of 82,221 words each were compiled to assess the priming effects of NSM. One was drawn from published papers on NSM and one from control linguistics publications by the same authors. Though the texts varied in length, the same number of words from each author was included in each corpus, to facilitate the comparison of raw totals and percentages in the two corpora. For example, the NSM corpus included Michaelis’ (2004) paper ‘NSM and cognitive-functional models of grammar’. The non-NSM text by Michaelis (2006) that was nearest in length to this paper was ‘What is an action-based model of interpretation?’ Once the reference list was deleted from these two papers, the NSM paper was 2,049 words and the non-NSM paper was 2,293 words. To render the papers equivalent in length, the first 244 words were deleted from the main text of the non-NSM paper, and the remaining 2,049 words from this paper were added to the control corpus. The full 2,049 words of the NSM paper were added to the NSM corpus. All papers that were reduced in length were deleted from the introduction, beginning after the abstract. Some of the NSM
papers did not discuss NSM explicitly until later in the paper, so the start of
the papers was deemed to be less relevant than the end. In all, fifteen authors’
works were included in the two corpora.

In each 82,221-word corpus, all explications in NSM were excluded.
NSM ‘explications’ consist of texts written using only NSM primitives,
which are employed to describe word meanings that NSM theorists judge
to be more complex than the meanings of the primitives themselves. For
example, Wierzbicka (1985) defines threat with the following explication:

if you do X I will do something bad (Y) to you
I say this because I want to cause you not to do X
(Wierzbicka, 1985: 495)

NSM explanations in the corpus were ignored in this study. Other excluded
words include primitives cited in the text, all example sentences, English
glosses of non-English data, titles of works by other authors, and all other
material in quotes, italics, capitals or small capitals, both in the NSM and
control corpora. For instance, when Travis (2004: 262) writes ‘I would argue
that the sense of sympathy in these examples is derivable from pobre “poor
thing” and from the context’, the words poor and thing were eliminated.
When Wierzbicka (1985: 502) writes, ‘the speaker says something “bad”…’
the word bad was eliminated, although the primitives says and something
were not in quotes and so were counted in the study. In the NSM corpus,
16 percent of the 82,221 words in the corpus were eliminated from further
study according to the procedures described above, leaving 68,955 words. In
the control corpus, 13 percent of the 82,221 words were eliminated, leaving
71,843 words.

A list of semantic primitives was prepared, based on the sixty-five
semantic primitives recently proposed in NSM (Goddard, 2008). Primitives
having more than 150,000 hits in the British National Corpus were eliminated
from the study. These words were not included for two reasons. First, the
high frequency of the sequences of letters that form words such as one, do
or the pronoun I would make it difficult to distinguish these words from
non-word sequences in the corpus, multiplying the complexity of the data
collection task. Exclusion of these high-frequency words therefore simplified
data collection. Second, words that are more frequently encountered in
a language are less likely to give rise to repetition priming effects than
infrequent vocabulary, so these words were regarded as less of a loss to
the study than their more infrequent counterparts (Becker, 1979; Forster
and Davis, 1984; Rubenstein et al., 1970; and Versace and Nevers, 2003).
Eliminated words included fifteen of the sixty-five primitives: I, you, this,
one, two, some, all, do, have, when, time, not, can, if and more.

The remaining fifty primitives from Goddard’s list were someone, people,
person, something, thing, body, the same, other, many, much, good,
bad, big, small, think, know, want, feel, see, hear, say, word, true, happen,
move, there is, there are, live, die, now, before, after, a long time, a short
If you study a word do you use it more often?

*time, for some time, where, place, here, above, below, far, near, side, inside, maybe, because, very, kind of, part of and like. Note that eight of the primitives contain two or three words. These phrases were searched for as units. That is, ‘of the part that I read’ would not be considered to contain a token of the primitive part of, even though it includes the words part and of. For convenience, all fifty words and phrases will be referred to here as ‘primitives’, though they are technically only primitives when they are used in NSM explications and not when they are used in English.

The two corpora were searched for the fifty primitives listed above, and the number of tokens was recorded. All inflected forms of the primitives were considered. For instance, says, said and progressive saying were considered alongside say. However, stems with derivational morphology were not counted. For example, placement was not counted as an instance of place. The extent to which place can prime placement, for example, is not known, so conservatism was called for in selecting the words that were included in the study.

For each text included in the corpus, the number of primitives in that text was rendered as a percentage of the total words from the text. To evaluate the differences in the frequency of primitives between the two corpora, a model selection process was employed, beginning with a six-way ANOVA including the categorical factors ‘nsm’ (with the levels ‘NSM’ and ‘control’), ‘author’ (with fifteen levels corresponding to the fifteen authors in the study) and the continuous predictor ‘year’ (that is, year of publication). The variable ‘author’ was nested within ‘gender’, ‘native language’ (either English or non-English, based on country of childhood residence) and ‘stance towards NSM’ (either supportive or critical). ‘Stance towards NSM’ was based on an assessment of the concluding or final sections of the papers in the corpus, in which the NSM approach was declared ‘valuable’, ‘effective’, ‘precise’ or ‘successful’ by the four authors here categorised as supporters of NSM (Goddard, 1997: 207; Stock, 2008: 753; Travis, 2004: 269; and Wierzbicka, 1985: 511), or else was criticised as, for example, ‘idiomatic’ or even ‘decidedly odd’ by the eleven authors here considered to be critical of NSM (Geurts, 2004: 225; and Kay, 2004: 241).

Non-significant interactions and factors were sequentially eliminated beginning with the largest p-values, resulting in the minimal adequate model.

3. Results and discussion

The fifty primitives comprised 3.31 percent of the 68,955 words in the NSM texts and 2.40 percent of the 71,843 words in the control texts. Potential factors and interactions influencing the percentage of primitives were evaluated in a model selection process that ultimately resulted in a monofactorial ANOVA with only ‘nsm’ (that is, whether or not the papers dealt with NSM) as a significant main effect (F≈8.6; p < 0.01), as shown in
Figure 1: Bar plot comparing percentage frequencies of the two levels of factor ‘nsm’

Figure 1. This effect of ‘nsm’ was relatively large ($\eta^2 \approx 0.24$). Variances of the variable values were homogenous according to Levene’s Test ($p \approx 0.54$) and residuals were normally distributed ($p \approx 0.15$).

The factor ‘nsm’, reflecting the difference between the NSM corpus and the control, was the only variable that was found to be significant. No other factors significantly affected the percentage of primitives appearing in the texts.

The factor ‘stance towards NSM’ was the last to be eliminated in the model selection process. NSM theorists have presumably been working longer and more intensively within the theory than the critics of NSM included in the corpus. This longer history of exposure to the primitives might be expected to lead to increased priming effects for the pro-NSM theorists (Bock and Griffin, 2000; Chang et al., 2000; and Kaschak et al., 2006). However, the pro-NSM theorists did not use significantly more primitives than anti-NSM theorists, either when writing about NSM or other semantic topics. This suggests that only the shorter-term effects of working on an NSM or non-NSM topic influenced the frequency of primitives in the researchers’ texts.

The factor ‘year’ might also have been expected to show an effect, given the fluctuating status of NSM over the period covered in the corpus, and the greater exposure to NSM over time by its proponents. However, priming by NSM primitives has not changed over the thirty-five years from which data was included in the corpus.

In sum, the NSM corpus contained significantly higher counts of the fifty primitives examined in the study. This suggests that studying and
If you study a word do you use it more often?

writing about NSM can lead to a greater use of the primitives employed in NSM analyses. Other potential influencing factors, such as authorial position on NSM, author gender, native language and year of publication did not significantly affect the number of primitives employed.

4. Conclusion

This study suggests that lexical repetition priming, like syntactic priming, may be apparent in a corpus. Corpus studies can provide a valuable avenue to examine the effects of lexical priming on linguistic analyses, given that corpus studies can target populations, such as professional linguists, which would be problematic to sample in a psycholinguistic experiment. These studies can also consider longitudinal data, such as the thirty-five years of texts considered in the current study, which would require substantial effort and planning to acquire in laboratory studies.

Due to its examination of linguistic texts, this study has relevance specifically for the field of linguistics. It is generally recognised that introspection, or linguists’ reliance on their own intuitive judgements, remains a dominant methodology in linguistics (Boroditsky, 2000; Gibbs, 2006; Glucksberg, 2001; Murphy, 1996; Schütze, 1996; and Veraeke and Kennedy, 1996). Linguists who rely on their own judgements as native speakers assume that they are, in fact, representative speakers of a language. Syntactic judgements have previously been shown to be influenced by factors including linguistic training and repetition priming (Dąbrowska 2012; Nagata, 1987a, 1987b, 1988, 1989; Schütze, 1996; Snyder, 1994; and Spencer, 1973). This study shows that linguists are also subject to lexical repetition priming.

To assess comprehensively the effects of repetition priming on linguists’ work, this study must be expanded in several directions. First, it will be necessary to compare the effects of repetition priming on linguists’ writing with the effects on their speech. Short-term effects on linguists’ speech are predicted to resemble effects on writing, based on observed similarities between repetition priming in these two modalities (Branigan et al., 1999; Cleland and Pickering, 2006; Hartsuiker and Westenberg, 2000; Kaschak, 2007; and Pickering and Branigan, 1998). Second, it would be advantageous to compare linguists and non-linguists to evaluate whether degree of linguistic training facilitates or hinders lexical repetition priming. This comparison would be most straightforward in a laboratory study in which advanced students of linguistics are compared with students who have not studied linguistics, along the lines of syntactic priming studies (Spencer, 1973). Third, the persistence of lexical repetition effects over time requires further examination. The NSM theorists in the current study did not demonstrate significantly greater priming effects than the non-NSM researchers, suggesting that the NSM theorists’ longer and more intensive exposure to NSM was less relevant than the shorter-term exposure involved.
in writing a research paper. Repetition effects have been demonstrated to last half an hour (Hughes, 2002) to an hour (Albrecht, 2008) after the time of exposure to words, but are hypothesised to persist after a longer delay (Bock and Griffin, 2000; and Chang et al., 2000). Further study is, therefore, required to identify the length of time over which lexical repetition priming can influence production.

To summarise, it is evident that the atypicality of linguists’ intuitions should be taken into account in linguistic analyses of the lexicon. The existence of lexical repetition priming suggests that introspection-centred methods are suspect in lexical studies as well as in syntactic studies, and that alternative methods of data collection such as corpus or experimental studies may be more effective in studies of both syntax and lexicon.

References


If you study a word do you use it more often? 285


If you study a word do you use it more often? 287


Texts included in the corpus


If you study a word do you use it more often?


